SMU Data Science Review

Volume 3 | Number 2 Article 15

2020

Cover Song Identification - A Novel Stem-Based Approach to Improve Song-To-Song Similarity Measurements

Lavonnia Newman
Southern Methodist University, lavonnian@smu.edu

Dhyan Shah
Southern Methodist University, dhyans@smu.edu

Chandler Vaughn
Southern Methodist University, cvaughn@smu.edu

Faizan Javed
Southern Methodist University, fjaved@smu.edu

Cover Song Identification - A Novel Stem-Based Approach to Improve Song-To-Song Similarity Measurements

Lavonnia Newman[1], Dhyan Shah[1], Chandler Vaughn[1], and Faizan Javed[2]

[1] Master of Science in Data Science, Southern Methodist University,
[2] Dallas, TX 75205, USA
[3] Adjunct Lecturer, Data Science, Southern Methodist University,
[4] Dallas, TX 75205, USA
[5] lavonnian, dhyans, cvaughn, fjaved@smu.edu

Abstract. Music is incorporated into our daily lives whether intentional or unintentional. It evokes responses and behavior so much so there is an entire study dedicated to the psychology of music. Music creates the mood for dancing, exercising, creative thought or even relaxation. It is a powerful tool that can be used in various venues and through advertisements to influence and guide human reactions. Music is also often "borrowed" in the industry today. The practices of sampling and remixing music in the digital age have made cover song identification an active area of research. While most of this research is focused on search and recommendation systems, plagiarism is a real industry wide problem for artists today. Our research seeks to describe a framework of feature analysis to improve cross-similarity, song-to-song, similarity distance measurements. We do this with the context that cover songs represent a fertile training ground to prove methods that can later be applied to plagiarism use cases. Our proposed method preprocesses songs by first source separating the songs into its constituent tracks prior to feature generation. This is otherwise known as "stemming". These subsequent spectral and distance features are then analyzed to provide evidence of improvement in overall modeling and detection performance. We find that using stem distances and overall distance measures achieves an uplift of 61.8% increase in Accuracy, a 59.2% increase in AUC, a 304.7% increase in Precision, and a 105.1% increase in F1 score for a regularized logistic regression. This process can be directly applied to more sophisticated deep learning frameworks.

1 Introduction

Most people you speak with love music. It evokes emotion and enriches our lives. In many ways it is a core component to our human existence, and it is embedded

in our daily lives so much that it can often go unnoticed. Unintentional music exposure occurs through commercial advertisements, elevators, bars, restaurants, stores, trains, planes, movies, sporting events and many more venues too numerous to list. It is used to set the mood for the occasion such as dancing, aerobic exercise, praise and worship, or even relaxation or work. When combining intentional music listening with background music, the average American is exposed to music more than 30 percent of their waking hours per day [1].

Music is also often "borrowed" in the industry today. Sometimes this borrowing happens as an approved and royalty giving business arrangement with the original artist, sometimes this borrowing happens by a direct copyright violation, and sometimes the borrowing happens inadvertently. When this practice only takes a beat line, or melody component, or some other part of a song, this is known as "sampling" the song. These practices have created a difficult environment for artists and music industry labels in protecting their work, and as a result has created a relatively healthy litigation environment in the music industry today. Further, individual musical tastes vary considerably. And, in this day and age, anyone with a few thousand dollars can produce an album in their garage and distribute it via the Internet. This, coupled with considerable business model changes for the music industry towards streaming, have created a rapidly advancing need for automation and intelligence in categorization and analysis of musical works that exceed the old standards of musical analysis.

An area of research that has resulted from this environment is something affectionately known as "cover song identification [2]. " A "cover song" is officially classified as a song that is re-performed and re-recorded by a separate artist. Often this is done with different instruments, different mixing, different tempo, or even a different gender singer singing vocals. It can be said that cover songs share partial characteristics of the original song. This can include melody and lyrics from the source song, but can be modified to include new musical elements, language, or rhythm and beat to suit a modern day taste. Identifying cover songs emerged as a computer science and data science problem as a result of the prolific digitalization of music. Cover Song Identification represents an algorithmic way to identify when two songs are derived from the same source composition. As might be obvious, a cover song can seem highly divergent from the original recording, which make identification a difficult opportunity.

Cover Song Identification is usually a computational query-retrieval task. Most music forensics analysis tools have their roots in signal processing, but many of the practices are still nascent and evolving. A forensic musicologist describes someone that analyzes music not as a singer, composer or writer, but as someone that is focused on analyzing and detecting structural similarity in music [3]. The breadth of research yet devoted to music similarity measurement is relatively finite. This is especially true when considering the proliferation and acceleration of adoption of digital music, and the increasing practice of sampling during the creative process for new works. Zhang, et al. classified research areas in

music similarity as generally belonging to the categories of "Metadata-based similarity, content-based similarity, or semantic description-based similarity" [5]. Given the nascent field, we chose to restrict our research discussions to content-based methods. This rich area provides a rich feature playground, with low-level, but high dimension features to work from.

Content-base similarity approaches include studies on estimating and matching signaling features such as chroma, harmonic pitch class profiles (HPCP), and mel-frequency cepstral coefficients (MFCC). These approaches and song fingerprinting can be used to identify cover songs, a related problem to the aforementioned sampling problem. However, few approaches specifically target building algorithmic approaches to detecting music similarity, or sampling as "a song within [another] song" [4][5]. Our research seeks to describe a framework of feature analysis to improve cross-similarity, song-to-song, similarity distance measures by first "stemming" the song into its constituent parts. Through this approach, we show that by considering comparison distance measurements of the full song pair, in addition to the stemmed and paired components for the drum, bass, vocal, and other stems, we can improve cover song identification processes. Potentially these methods can eventually be applied to detect song plagiarism directly, where plagiarism may include only a portion of a song, which we also explore [6].

The remainder of this paper is organized as follows:

- Section 2 provides an overview of how music is currently analyzed, and the associated properties that are measured,
- Section 3 covers ethics and ethical implications for this work,
- Section 4 provides a survey of techniques that are important to understand
- Section 5 covers our proposed framework and approach for cover song identification and how it may apply to plagiarism,
- Section 6 describes the empirical results to our testing and experimentation,
- Section 7 suggests future avenues for study,
- Section 8 outlines lessons learned from this research, and;
- Section 9 concludes our findings.

2 Music Analysis and Properties

To understand how we intend to analyze audio for similarity, its important for us to take a moment to educate the reader to the feature-rich space that audio provides. While the terms "unstructured data" typically elicits ideas of numerous data points or text attributes, on several topics, with no obvious way to statistically analyze them, in this case we use it to definitively describe raw audio. While we will not attempt to cover the full sphere of audio analysis, we

provide the overview below on audio analysis as a baseline starting point for the reader to further understand our work, and to appreciate the complexities of audio analysis.

2.1 Overview – Background Definitions

Most readers will relate to common audio concepts such as frequency and amplitude. Frequency represents the speed of a vibration, which thus determines pitch. Amplitude, in contrast, is the size of that vibration. So if frequency determines pitch, amplitude can be thought of deterministic to how loud that pitch is. The spectrum of human-detectable hearing ranges to 12 distinct pitches, through various octaves. This, at its core, makes up the basis of a sort of DNA for music and coupled with temporal, rhythmic and melodic features, it is the basis of modern western music that we know today. This also provides a basic framework for understanding some of the audio feature extraction techniques we describe below.

One principle area of audio feature extraction centers on a term described as chroma. Chroma is representative of tonal and pitch content. In literal terms, it is the "color" of a musical pitch, decomposed such that it is octave-invariant into 12 pitch classes. Chroma features help capture musical characteristics in a condensed matrix, and potentially visual, form while "being robust to changes in timbre and instrumentation" [7]. Chroma features are typically extracted from raw audio content utilizing Short Time Fourier Transforms, Constant Q Transforms, and normalized Chroma Energy methods, as shown in the following (Fig. 1) example processing pipeline.

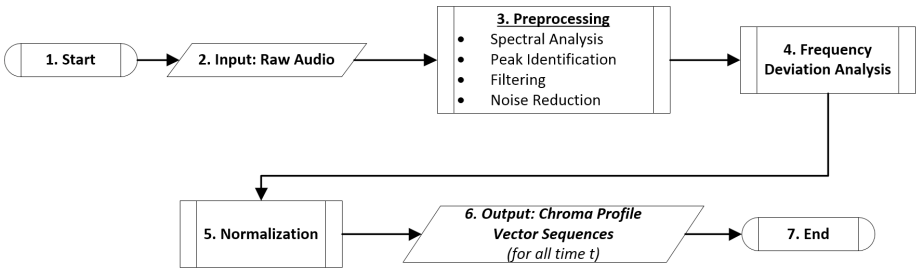


Fig. 1. Chroma Feature Extraction Process

It is logical to think that for any time t in a local time window of a song, it would have distinct chroma features. If we think of building these chroma feature windows across all windows for the song, we can build a representative pitch profile of the song over all 12 pitch chroma bands. When combined, this creates

what is known as a chromagram for the song (see Fig. 2. Chromagram Example). Chromagrams are sometimes displayed as the squared magnitude of the Fourier coefficients at each section. In these cases, it is known as a spectogram. This spectogram is one way to densely represent a song's pitch structure.

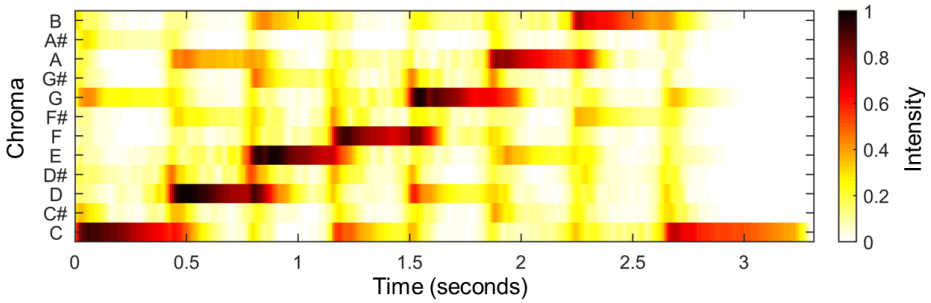


Fig. 2. Chromagram Example

Chromagrams can be thought of as one of the many building block features to audio processing. We cover it here only to instruct the reader on the type of processing that is done in order to extract features from audio. Several different preprocessing and post-processing techniques can be applied to raw audio to yield varying spectral, rhythmic, temporal, and melodic features for raw audio. Some of these subsequent features are tabulated below. While an exhaustive review of how these features are calculated is beyond the scope of this paper, we encourage the reader to review references for these feature types to learn more. For those features we utilize directly, we will explain in line with their use.

Some of the more common tools used today for music and speech recognition include mel-frequency cepstral coefficients (MFCC), and spectrogram analysis. Unfortunately, both of these methods are sensitive to additive noise and pitch dynamics. As a result, it is common to see log scaling as a post-processing step. (see Fig. 3. Spectogram Log Scale Example and Fig. 4. Mel-frequency Cepstral Coefficient (MFCC) Example)

2.2 Feature Extraction

Now that we have covered some of the definitions and concepts around the chromagram extraction process, let's walk through the process leading to feature extraction that will ultimately allow us to detect similarities between two musical compositions. As shown in Fig. 5. below the process begins with the music file, specifically a WAV file. While one might think the file format does not matter, it actually does. The various music file formats are used for particular purposes.

Table 1. Example Acoustic Features

Category	Category Description	Feature Space
Timbral	Tonal texture	Harmonic Pitch-Class Profile
		Spectral Centroid
		Spectral Contrast
		Rolloff
		Low-Energy
		Mel-frequency Cepstral Coefficient
Temporal	Time domain signals	Zero Crossing Rate
		Autocorrelation
		Waveform Moments
		Amplitude Modulation (loudness)
Spectral	Musical characteristics by spectra	Auto-regressive features
		Spectral asymmetry
		Kurtosis
		Flatness
		Crest factors
		Slope
		Decrease
		Variation
		Frequency derivative of Constant-Q
		Octave-band signal intensities
Rhythmic	Musical timing	Beat histogram
		Rhythm strength
		Regularity
		Average tempo
Melodic	Melodic content	Pitch histogram

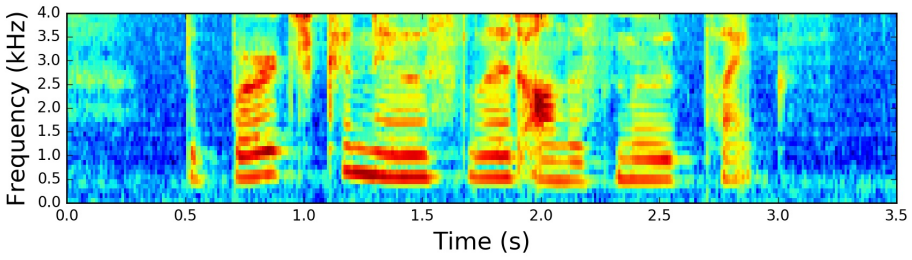


Fig. 3. Spectogram Log Scale Example

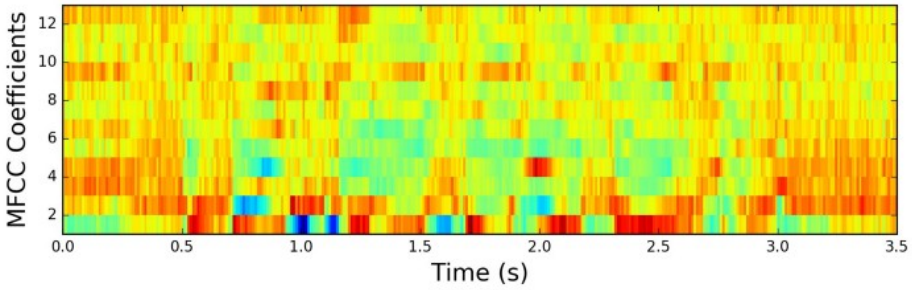


Fig. 4. Mel-frequency Cepstral Coefficient (MFCC) Example

The most popular file type MP3 is used for the distribution of music, downloads and posting on websites. It is compressed, uses little disk space, and offers a CD like quality that is appropriate for listening and multimedia presentations. MP3's adoption was fueled by the adoption of the Internet due to its compression and quality profile. A WAV file, by contrast, is uncompressed, with no quality reduction induced, and is much larger in size comparatively. WAV files provide higher quality sound than MP3 and are better for music analysis and editing as a result. Additionally, WAV files can be looped together for repetition and can provide a seamless playback while MP3 files cannot. The first 10ms to 50ms of a MP3 snippet will always have a dead space that is due to the compression algorithm that created it. You cannot loop a MP3 sound bite without having those silent gaps. If comparing the concept of an audio WAV file to the video presentation, the WAV file would be considered high resolution while MP3 would be considered standard resolution. Even though WAV file format is preferred for music analysis, we can achieve relevant results by processing MP3 files as well.

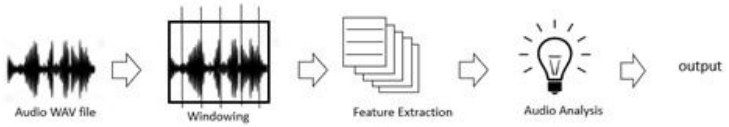


Fig. 5. Audio Analysis Process

Numerous processing frameworks exist for processing raw audio and extracting both high-level and low-level features. We selected Essentia for analysis and feature extraction of audio features as we felt, after review, it represented the current state of the art in wide distribution. Essentia is an open source library that was developed in C++, with Python wrappers, for the purpose of audio signal processing and audio music information retrieval. Essentia was used to process audio and divide it into short audio segments. This process, which is

called windowing, results in a continuous sequence of blocks that can ultimately be subjected to audio signal analysis. The window function length is configurable and often is based on the length of a particular event or feature of interest [9].

After segmenting the music stream into audio blocks, features are extracted. Features, which are the contextual data of an audio signal, are the foundation of research and development in audio signal processing. Features capture the physical and perceptual impact of the signal. The goal of feature extraction is to obtain the notable characteristics of a signal and converting them to vectors or matrices of numerical coefficients. Once extracted these features can be used for data mining, data classification, statistical analysis and in our case similarity measurement. In the following paragraphs we provide insight into the various descriptors and the associated features used for music information retrieval that form the analytical data for this project.

- **Timbral Descriptors** - The features in this domain are related to the timbre of the music or the tone quality. Tonality in music is the organization of the note on a musical scale. It describes the sound's structure that is composed of harmonized related frequencies. The brightness feature would also be contained in this domain.
- **Temporal Descriptors** - These features are unique in that they do not require the audio signal to be transformed. These computations occur on the original signal sample. Overall the temporal domain features capture the duration of the signals, the loudness, and durations that cross a peak energy level. Common terms for capturing the loudness and energy levels consist of *amplitude*, which is the temporal structure of the signal, and *power* which is also temporal but focused on the signal's power. One of the popular features in this category is the zero-crossing rate, which is the rate the signal goes from negative to zero to positive amplitude and vice versa, which is often using in classifying percussive sounds.
- **Spectral Descriptors** – These features are also known as frequency-based features. This domain contains the largest group of audio features. The features are derived from an autoregressive analysis, or a Short Term Fourier Transform (STFT), and describe the physical properties of the signal frequency. Autoregressive features capture the results from a linear prediction analysis of the signal while the STFT captures a derivative from the signal's spectrogram. One of the benefits of STFT is that it is able to capture how the signal changes through time, which has made it a foundational component to audio signal processing.
- **Rhythm Descriptors** - The rhythm features capture the organization of sonic events along a time axis [8]. The beats per minute, the beat loudness, and other beat tracking features are found in this domain.
- **Melodic Descriptors** – The features in this category are related to the melody, pitch, chords, harmony, and tuning.

We hope given this high-level overview of audio processing, we have primed both interest and intellectual curiosity for audio processing. For this paper moving forward, we will focus specifically on the following features for analysis:

- **Mel-Frequency Cepstral Coefficients (MFCC)** - Derived from the cepstral representation, these coefficients are equally spaced into bands in an effort to better approximate human capabilities for hearing [8][9].
- **Gamma-tone-frequency Cepstral Coefficients (GFCC)** - "This is an equivalent of MFCCs, but using a gammatone filterbank scaled on an Equivalent Rectangular Bandwidth (ERB) scale." [9]
- **Harmonic Pitch Class Profile (HPCP)** - "HPCP is a k*12 dimensional vector which represents the intensities of the twelve (k==1) semitone pitch classes (corresponding to notes from A to G#), or subdivisions of these." [9]
- **Chroma Cross Simularity (CSS)** - This feature is the combine binary similarity matrix for two different chromagrams from two different songs. The CSS is dependent on the HPCP as a precursor. Or said differently, the CSS is calculated by comparing two HPCP's [9].
- **Cover Song Similarity** - An algorithm for computing the Smith-Waterman score matrix between two chroma cross similarity matrices [9][10].

Once features are extracted, files containing the associated values were generated for each of the signal chunks. These files can be used as the data input files to be merged and utilized for exploratory data analysis prior to engaging in any modeling.

2.3 Dynamic Time Warping

Dynamic Time Warping (DTW) can be used to compare the rhythmic patterns between two songs. DTW calculates the non-linear distance between the original and infringed song's feature sets. The smaller the distance the closer aligned the features. DTW is unlike the Euclidean distance measure in that it warps the frames in a non-linear fashion together accounting for time and speed differences (tempo) of the songs. In this way DTW is robust to tempo differences between songs. Pictorial representations of the mappings are shown in Fig. 6 and Fig. 7. In prior work researched by Ren, Fang, and Ming fusing GFCC and MFCC features together prior to calculating the distance using DTW, yields an improvement in the recognition rates by 21.3 percent over the non-fused features [11]. The following figure is a pictorial comparison of the Euclidean and DTW distance algorithm [12]. We utilize both Euclidean and DTW distances to achieve our results.

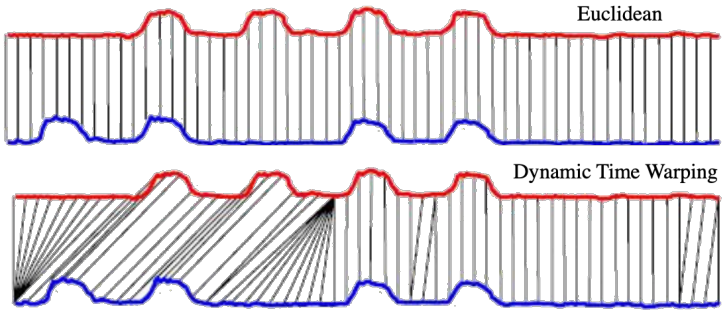


Fig. 6. DTW and Euclidean Mappings

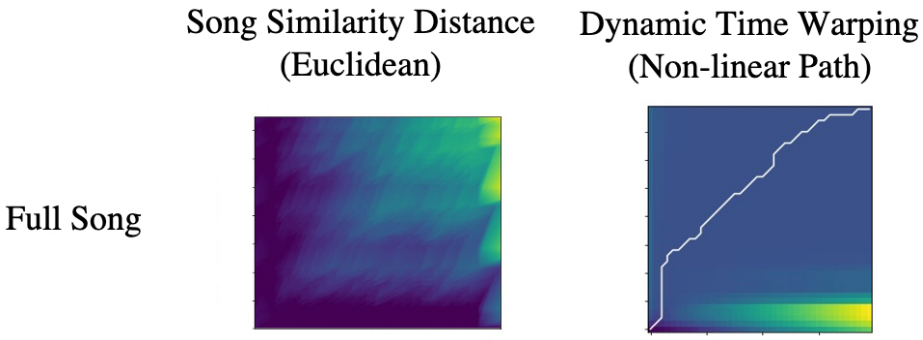


Fig. 7. Full Song Euclidean and DTW Distance Mappings

3 Ethics

There are several ethical considerations in this analysis. Obviously in music there are the general concerns over plagiarism and copyright infringement. A use or close imitation of a substantial part of another musician's work claiming one's own music without a proper credit to original artist constitutes music plagiarism. Further, the music industry attempts to restrict classifications of "sound-a-likes," making it difficult for anyone to make a claim that their song sounds like another artist. These issues complicate technological evolution since finding music similarity is, by its definition, the process of finding songs that sound like other songs.

Further, from a researcher perspective, there are ethical considerations around bias. Researchers may tend to choose known or readily available algorithms and constructs to leverage in their research, which may inherently bias music similarity to the potential end detriment to certain artist's music. Plus, music is still very much a human experience. Emotion, mood, and feeling are a continuum

and not necessary ordinal or categorical measurements. This also has direct implications for ethical considerations [7].

3.1 Music Industry Ethical Considerations

Music similarity is only considered plagiarism if the original work has copyright protection. However, plagiarism does not necessarily result in a copyright infringement. Worldwide, the music industry is suffering from what might be considered as plagiarism, as end-users have instantaneous, easy access to digital media and contents over the internet, and can easily sample, re-sample, and modify existing artistic works by incorporating them into new derivative works.

There are few defined standards to evaluate music similarity or plagiarism, and the intellectual property laws in the U.S. fall short of clearly and legally defining the issue. The most recent Digital Millennium Copyright Act defines the statute of limitation to 70 years after the death of the creator of a musical work, which has also blurred the lines between the original song and an allegedly infringed song [7].

Venturing into music forensics and applying data science techniques, one must be cautious in evaluating music similarity. Ethically speaking, there can be three possible categories for musical similarity [13]:

1. Inspiration,
2. Coincidence, and;
3. Plagiarism, copying, wrongful appropriation.

Inspiration is considered a legitimate element of similarity in any form of artistic production. For centuries musicians have been genuinely inspired from other variants of music (e.g. folk music). As long as inspiration is limited to only fragments of artistic bites and synthesized into a musician's own perspective to create original music, this is usually culturally acceptable.

Additionally, there are only 12 notes in Western music and therefore coincidence of musical similarity between two songs have a true non-zero probability. Thankfully, there are many multifaceted dimensions to a song. In legal terms however, as the dimensions of song similarity increase its less likely to be considered coincidence. Today most of the legal industry operates under subjective measures, rooted in musicology practices. These measures are increasingly being routinized over time, but in no time soon will it be complete.

When taken together in the context of music similarity, the conditions of the first two categories help prove the third. If there is the potential for inspiration, and the probability for coincidence is low, then there is a high likelihood that

plagiarism has occurred. This is when a song potentially violates copyright law. The prevalence of sampling in pop music today makes this a particularly difficult and widespread issue in the music industry today [14]. Where does inspiration start and stop? How different can a rhythm or melody be to be original? These are questions that the entire industry is grappling with.

3.2 Research Ethical Considerations

In the area of bias as it relates to research, the questions are different. There are certain considerations we can be safe to consider moot. For example, the idea that music similarity research results could cause physical harm to listeners of the music is not plausible under normal circumstance. However, could this research be used to economically harm certain groups or individuals? Could it disadvantage an artist or group of artists as a result? Unfortunately, without proper controls and forethought the answer would be "yes" there is a probability that benefits to some could happen at the expense of others.

Unexpected or undue harm can occur as a result of "unintentional power and bias" dynamics in the research process itself [7]. Due to the complexities of algorithms, and the embedded nature of this intelligence into products, it creates a *black-box* problem we have to take into account. Holzapfel, et. al [13] claim that bias come in three forms:

- **Pre-existing** – relating to existing socio-cultural norms.
- **Technical** – relating to the data available, methods, or evaluation techniques.
- **Emergent** – relating to how algorithms behave when faced with new emergent types of data that they have never encountered before.

As researchers, we would also posit that we have our own form of research bias that we must consider. We obviously want to help solve a problem, and contribute to the global knowledge base with the quality of our work. This unfortunately creates potential bias to create a favorable looking body of work, which could motivate researchers to make certain choices about data sets, about problem scope, and about the technical approach which may inherently be biased. Each of these forms of bias have the propensity to potentially impact artists and other market participants.

To address these concerns, we carefully considered our research design, validating our choices along the way. As guidelines we evaluated these decisions through the following lens:

This is the lens upon which we hold true for this body of work.

Table 2. Bias Category and Challenge Questions

Bias Category	Challenge Questions
Preexisting - Cultural	Are we considering the proper breadth of data to represent cultural minority classes, under-represented groups, and emergent categories?
Technical	Does our data choice contain a representative sample for the problem scope? Is the quality of the data such that we can be reasonably assured that it will not materially impact the results?
Emergent	Have we designed tests to validate results against new types of data properly?
Researcher Judgement	Do we have proper controls to document data or technical related issues to combat error propagation? Have we properly documented "ground truth" measures?

4 Survey of Techniques

There is a variety of prior research completed on detecting music similarities. The bodies of research focus on three primary areas: 1) music style recognition, 2) genre categorization, or 3) plagiarism. Bogdanov, Serrà, Wack, Herrera and Serra demonstrate music similarity measurement by facilitating multi-media retrieval and focusing on seeking a suitable distance measurement between songs, based on predefined music feature space [15] . Li and Ogihara demonstrate acoustic-based features to retrieve music features which focused on similarity search and music sound emotion detection by applying Db8 wavelet filters and timbral features in order to generate compact music features [16].

The general approach for similarity search is defined by finding a Euclidean distance measurement between the feature space of two songs, and for emotion detection the typical approach has been to decompose multi-class classification into a Support Vector Machine (SVM) problem in order to train a system on extracted features. In the case of Logan and Salmon [17], they employ methods of comparing songs solely based on their audio content. This creates a signature for every song based on their K-means clustering of spectral features. And from a machine learning approach, Dannenberg, Thom, and Watson [18] illustrate how to build an effective style classifier by identifying 13 low-level features based on MIDI data, which is classified using Bayesian, linear and neural network algorithms. Finally, Kuo, Shan, Chiang and Lee [19] created a personalized content and emotion-based music filtering system that predicted the preference of new music by computing a weighted average of all ratings given by the peer group of

similar preferences for end-users. All of this prior art critically moved the state of the art with reproducible research forward, and is acknowledged as such.

Approaches to the problem of cover song similarity has been studied in past as well [11,15,16,17,18,19,20,21,22]. Where Logan and Salomon offer insight on the use of MFCC to define similarity, Tralie proposes fusion of MFCC and HPCP features to improve overall performance of cover song detection [17][20]. Additionally, Chang, Choe, and Lee utilize Dynamic Time Warping along with GFCC and MFCC fusion techniques to achieve better results than MFCC or GFCC alone [21]. While full replication of this work is beyond the scope of this paper, we take lessons from these approaches to propose a new approach prior to feature fusion that we believe can improve cover song similarity measurement, and thus outcomes for downstream modeling. Like the prior research listed, we focus on exploring chroma-based feature spaces, and the value that they lend given their condensed representation of audio features. However, we performed this analysis using stemmed features and a combination of DTW and Euclidean distance measurements.

5 Solution Approach

Given the scope of improving cover song similarity measurement, it is critical to plan each stage of the solution. This is important not only for the build out towards analyzing the data, but is also critical for fast iteration as new insights emerge that take research in new, interesting directions. It is with this in mind, and with the end goal of building reproducible research, that we outline our solution in this section.

5.1 Data Acquisition

As with any project of this type and ambition, we need to ensure we have enough high-quality raw data to work from. After reviewing various data sources for music data and metadata, the C80 cover song data (c80) was chosen [23]. The c80 dataset includes a collection of 80 tracks (songs), each performed by two separate artists. This archive represents our MP3-based cover song data, organized by cover song, which is later converted into over 30 GB of potential audio features. In addition to c80 providing an accessible data set, it also provides a standard data set to perform research on as many papers have leveraged it for MIREX Conference research over the course of years [24]. MIREX is a common music information retrieval conference frequented by researchers in audio processing.

We chose to perform raw feature extraction on the data ourselves, as opposed to utilizing any prior art or accessible source. This served the following purposes:

a) it maintained targeted control of the process, and b) it ensured reproducible results. As one might imagine, much hinges on feature generation when it comes to signal processing and handling raw data for these tasks.

5.2 Technical Approach

Feature extraction is critical for this research, and the pipeline for that extraction equally so. As we mentioned earlier in this paper, we set out to prove that stemming songs prior to calculating cross similarity matrices, can improve overall performance and resolution for those distance measurements. To test this, we present and define a novel pipeline approach to feature extraction, as shown in the Fig. 8 below.

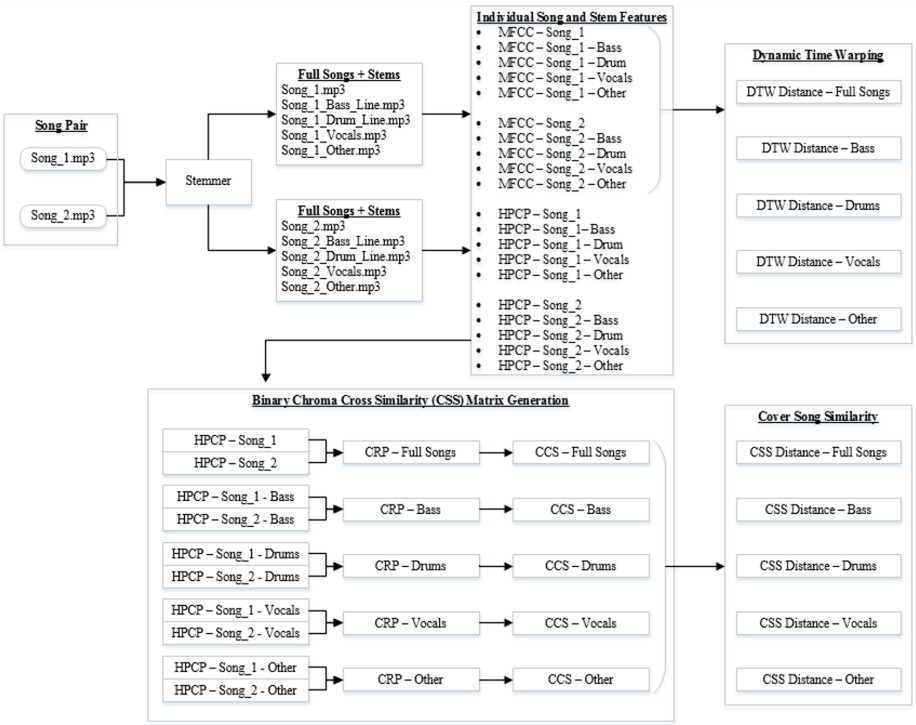


Fig. 8. Novel Stemming Pipeline for Cross Similarity Matrix and Distance Feature Generation. Features include: Mel-Frequency Cepstral Coefficients (MFCC), Gammatone-frequency Cepstral Coefficients (GFCC), Harmonic Pitch Class Profile (HPCP), Cross Recurrent Plot (CRP), Chroma Cross Similarity Matrix (CCS), Cover Song Similarity Distance (CSS), and Dynamic Time Warping Distance (DTW).

The body of potential audio features to explore can be dizzying. Essentia alone can generate 100's of potential features from a single song. As shown in Fig. 8, we first run the songs through a stemming process. We stem all songs into four constituent components. This procedure takes the full song and reverses the mixing of the song by splitting out the bass track, the drum and percussion track, a vocal track, and a catch-all "other" track. Thanks to recent advances in Deep Learning, there are existing models and tools that allow for this procedure with relative accuracy. While the intricacies of this stemming tooling is beyond our scope to explain in full, we encourage the reader to explore it. For our research we utilized the *Spleeter* package and associated models [25]. Spleeter was released to open-source in 2019, and was showcased at the 2019 ISMIR conference as the state of the art for source separation.

Features were generated for each song, stem, song pair, stem pair for each cover song pair, song to Gaussian white noise pair, song stem to Gaussian white noise stem pair, song to random song pair, and song stem to random song stem pair. In an effort to validate proper signaling and separation for the analysis every song Cover Song Similarity (CSS) was base-lined against a Gaussian white noise (0 dB) audio file, as well as a randomized song choice which was verifiably not identified as a direct cover song pair. This process was repeated for every song, which subsequently generated our feature space.

Our total data contains a total of 80 source covers, representing 160 songs, 640 song stems, and 4 Gaussian white noise stems. With added pairing for song to white noise comparisons, and randomized song comparisons, we had a rich feature space to work from. See Table 3 for the base-level observation breakdown. Total features processed are shown in the Table 4. As noted by Zhang, et. al., high dimensionality of existing [low-level] audio features has restricted the applicability of music retrieval and similarity calculations in large collections [5]. We take care to reduce this feature space through analysis and experimentation so that it is clear what effects are meaningful.

Table 3. Data Explanation

Data Origination	Quantity
Song Pairs	80
Songs	160
Song Stems	640
Guassian Whitenoise (0 dB)	1
Guassian Whitenoise (0 dB) Stems	4

Table 4. Feature Space Generated

Feature Type	Quantity
MFCC	805
GFCC	805
HPCP	805
CRP	1284
CCS	1284
CSS	1284

5.3 Exploratory Data Analysis

The covers80 data set is largely a binary and raw audio data set. Given this, exploratory data analysis (EDA) on the raw data is limited to reviewing wave forms, which is not meaningfully informative for this exercise. That being said, there are some significant observations we can cover before reviewing the feature space directly. Given the addition of white noise, stemming, and randomization to our data, our data will become highly imbalanced when viewed through the lens of determining whether a cover song is detected or not. For this analysis, the combined final base data set consists of 315 rows and 36 columns, with 24 of those columns as high dimensional arrays.

We purposefully ignore song metadata for this analysis since we are most concerned with determining the underlying similarity of audio wave forms. As such our EDA focused primarily on the generated feature space. Most informative is to review the HPCP's for the full songs, and the associated stems, to visually see similarities.

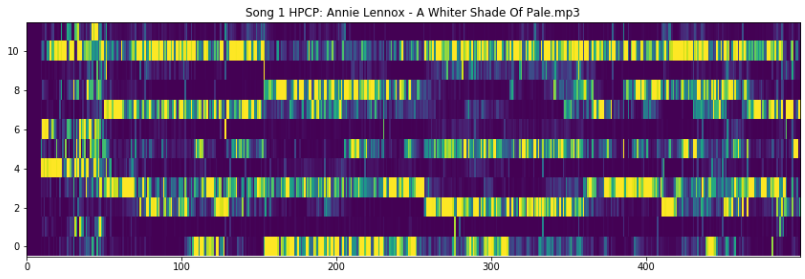


Fig. 9. Harmonic Pitch Class Profile - Original Song

Then we can review the mutual nearest neighbors binary cross similarity matrix. Longer interrupted diagonals in this graphic indicate a higher Smith Wa-

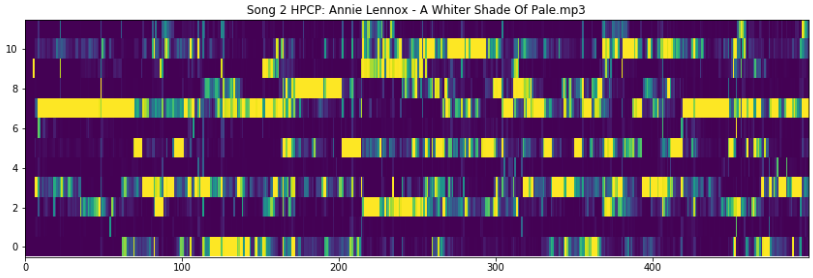


Fig. 10. Harmonic Pitch Class Profile - Cover Song

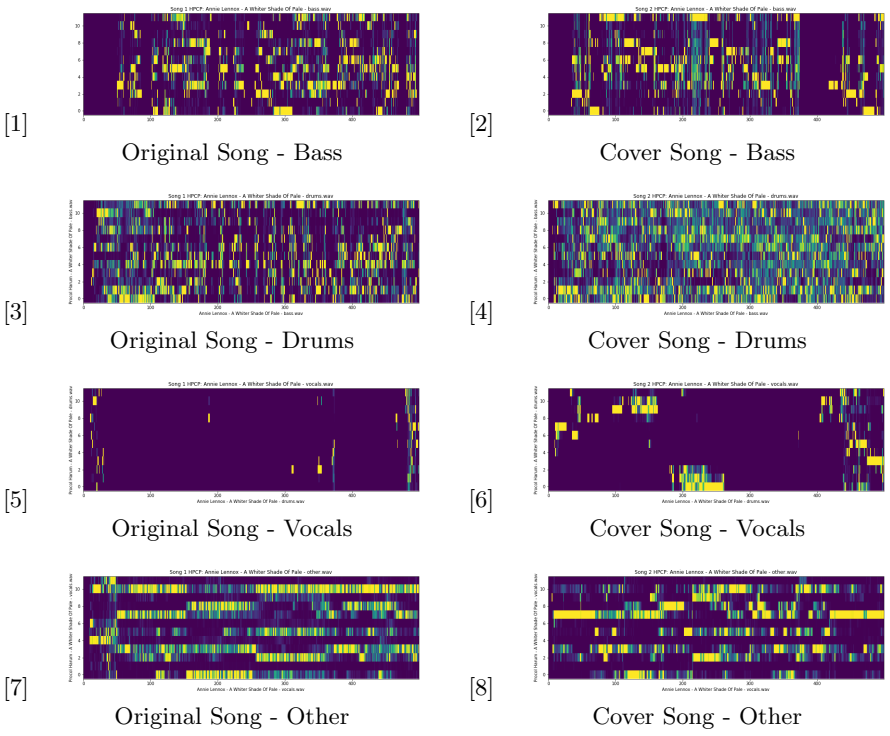


Fig. 11. HPCP Comparison For Cover Song Pair - A Whiter Shade Of Pale

terman asymmetric alignment, and thus yield a lower distance metric, as shown Fig. 12 ad Fig. 13 below.

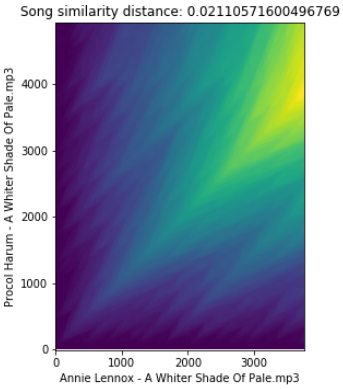
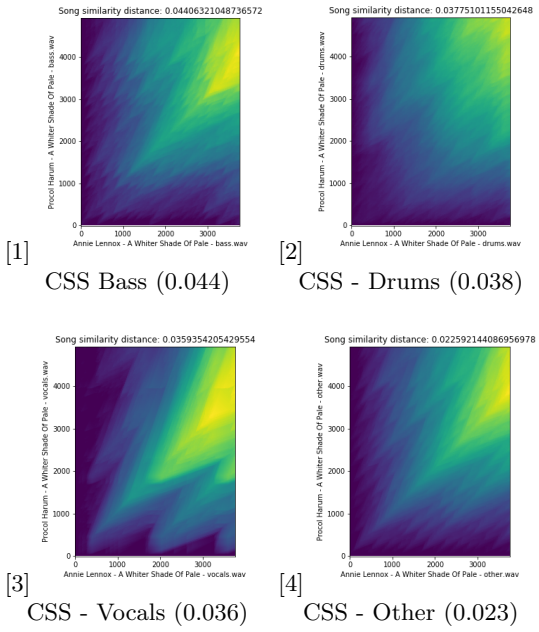


Fig. 12. Cover Song Similarity Matrix - A Whiter Shade Of Pale - Distance = 0.0211



[1] CSS Bass (0.044) [2] CSS - Drums (0.038)

[3] CSS - Vocals (0.036) [4] CSS - Other (0.023)

Fig. 13. Cover Song Similarity Matrix By Stem - A Whiter Shade Of Pale

In this example (Fig. 11, 12, and 13), for the *A Whiter Shade Of Pale* cover song performed by both artists *Annie Lennox* and *Procol Harum*, we see distinct long lines in the vocals and the other stems. The vocal similarity is to be expected given its a cover song. But what is striking with these visuals is the similarity between the overall similarity and the catchall *other* stem. Interestingly the overall distance score between these two songs is 0.021, and the distance between the other stems is 0.023, with bass, drums, and vocals stems with marked higher distance measures, 0.044, 0.038, and 0.036 respectively.

These similarity differences will become the foundation of our analysis. In order to determine whether we can see improvements in classifying cover songs, we reviewed the following assumptions:

- **Normality:** Initial review of the distributions show some variation between normality and spread for each of the distance measures. This is shown best in Fig. 14, which boxplots the pair distances in raw calculated form.
- **Standard Deviation:** Additionally the standard deviations appear to differ as well.
- **Linearity:** – All distances had a natural log transformation applied to ensure adequate spread and more homogeneous standard deviation. While some curvature exists with some variables, we accepted this and proceeded with caution.
- **Influential Values:** No legitimate reason existed to remove high leverage data points, so these were left in place.
- **Multi-collinearity:** With songs, and especially stemming, we must take care to handle multi-collinearity. We performed PCA on all data sets prior to estimation and modelling to reduce variable inflation factors (VIF). All distance values for VIF were <1.5 as a result, while maintaining a target of keeping $>99\%$ of information.

Having addressed assumptions we found this gave adequate spread in the measures, which is shown in Fig.15 and Fig.16. We make these aforementioned adjustments and proceed with caution.

6 Analysis and Results

We perform this analysis in four stages. First, we attempt to determine the efficacy of utilizing stems in cover song identification. This being a close analogue to plagiarism detection, it is our feeling that a stemmed approach to determining distance might better represent the changing market landscape of the music industry as it relates to track sampling and remixing of works. After determining efficacy, we analyze the impact of utilizing a minimum distance measure, Min(d),

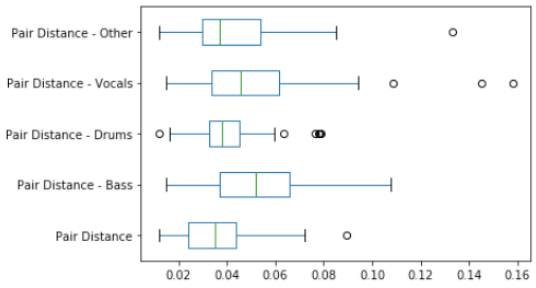


Fig. 14. Pair Distance Measures

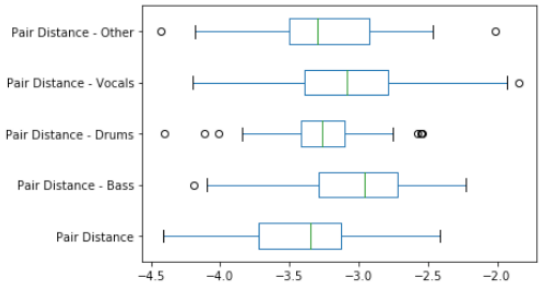


Fig. 15. Pair Distance Measures (Natural Log Transformed)

as a predictor for all Euclidean distance of d. We then determine the impact of using the fusion techniques with Euclidean and Dynamic Time Warping (DTW) distance measurements. We then demonstrate the usefulness in fusion methods with stemming.

6.1 Stemming Performance

In order to assess how stemming might impact cover song distance measurements we begin with determining whether the the distance measurements are meaningfully different from the overall similarity distance. We therefore define a new feature as the minimum of all distances $\min\{d_1, ..., d_n\}$ where n represents each distance measurement d for the full song pairs, and the subsequent stemmed pairs.

We first test whether the median of this new minimum distance measure is different than the median of the full song pair distance. Performing a one-way ANOVA between the median log-transform full song similarity difference versus the minimum log-transform distance of all measures, we define the following:

1. **Null hypothesis:**

$$H_0 : \eta_{fullsongs} = \eta_{min}$$

2. **Alternative hypothesis:**

$$H_A : \eta_{fullsongs} \neq \eta_{min}$$

Here η denotes the population median. We utilizes medians for inference since we are performing a one-way ANOVA on log-transformed data.

Table 5. ANOVA - Full Songs Distance vs Minimum of Distances

	Sum of Squares	**df**	**F**	**PR(>F)**
C(distance measures)	1.154747	1.0	7.339057	0.007522
Residual	23.916083	152.0	NaN	NaN

Decision: There is sufficient evidence at the $\alpha = 0.05$ level of significance ($p<0.007522$ from a one-way ANOVA) to suggest that the two medians are different between the distances for the full song, and the minimum distances found from all distances within the distance set for a cover song pair.

Given that there is evidence to suggest the medians are different, this lends credit to our hypothesis that we can utilize stemming to better signal whether a song pair is or is not a cover song. Plotting the log distances in Fig.16 we can see distribution shifts for song pairs that are cover songs, versus non-cover songs. Most notably, we can see the degree to which better separation is occurring by stem.

Performing a Tukey HSD post-hoc analysis also supports the use of stemming. Comparing the medians for all distance measure pairs provides evidence of value for modeling with Min(d) and stem measures. Or rather, there is sufficient evidence at the $\alpha = 0.05$ level of significance ($p<0.05$ from a Tukey post-hoc analysis) that the medians of Min(d) and the stems are different, as well as the median distance measures for the full song versus the median stem measures for bass and vocals.

Notably, this data runs the risk of multicollinearity, which can bring about instability in coefficients. Variable Inflation Factors (VIF) also indicted this to be of concern for several of our variables. To adjust for this prior to running any modeling, principal component analysis (PCA) was performed with a target of capturing greater than 99% of the variation. This effectively addressed the

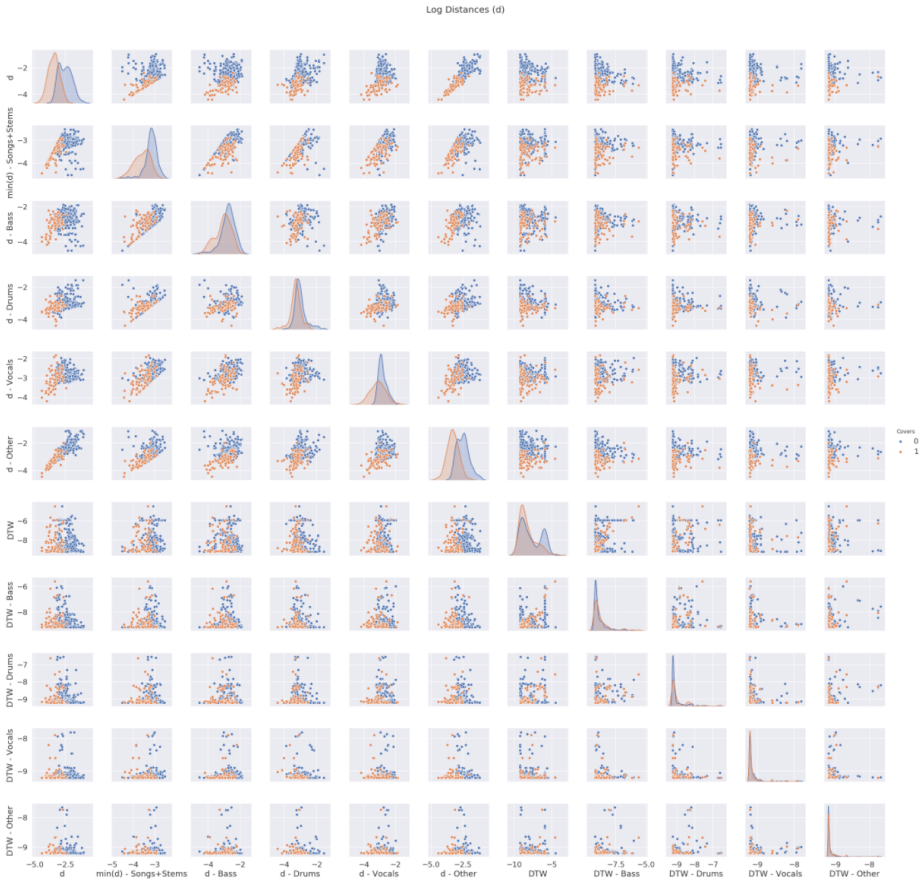


Fig. 16. Log Song Pair Distances - Distribution and Scatter Plot

multicollinearity concern and all factors subsequently had VIF values of less than 1.5, as noted previously.

In order to test how well stemming might work in the context of classification, we generalize this problem to a binary outcome as the most conservative case. Typically cover song similarity is considered a spectrum, which is why a lot of prior art represents results in mean average precision (MAP) terms. Here we choose to measure improvement for this binary classification in classical terms. We do this for clarity since the scope of this research is to prove that stemming is valuable as a preprocessing step, and that these techniques can be applied in a plagiarism detection context.

Utilizing our data we can test whether stemming improves prediction performance via a logistic regression. Keeping in mind we wanted to compare predictive

Table 6. Tukey HSD - Distance Measure Comparisons

Group 1	**Group 2**	**MedianDiff**	**p-adj**	**Lower**	**Upper**	**Reject**
Full Song Distance	Min(d)	-0.1732	0.1314	-0.3713	0.0249	False
Full Song Distance	d - Bass	0.3673	0.001	0.1693	0.5654	True
Full Song Distance	d - Drums	0.1455	0.3112	-0.0525	0.3436	False
Full Song Distance	d - Vocals	0.3249	0.001	0.1268	0.5229	True
Full Song Distance	d - Other	0.1624	0.1894	-0.0357	0.3605	False
Min(d)	d - Bass	0.5405	0.001	0.3424	0.7386	True
Min(d)	d - Drums	0.3187	0.001	0.1206	0.5168	True
Min(d)	d - Vocals	0.498	0.001	0.3000	0.6961	True
Min(d)	d - Other	0.3356	0.001	0.1375	0.5336	True
d - Bass	d - Drums	-0.2218	0.0169	-0.4199	0.0237	True
d - Bass	d - Other	-0.2049	0.0372	-0.403	0.0069	True
d - Bass	d - Vocals	-0.0425	0.900	-0.2405	0.1556	False
d - Drums	d - Other	0.0169	0.900	-0.1812	0.2149	False
d - Drums	d - Vocals	0.1793	0.1055	-0.0187	0.3774	False
d - Other	d - Vocals	0.1625	0.1889	-0.0356	0.3605	False

outcomes based on cross similarity between full songs, versus cross similarity for stems on song pairs, we performed logistic regression L2 regularization and cross validation iteration. We utilize a 70:30 sampling of the data for a training set and a test (hold out) set through each iteration. While intermediary models are too numerous for discussion here, the results of our test are shown in Table 7. Tests were run for models based on the full song cover song similarity score, just stemmed-based similarity scores, and with using stems and the full song scores together. PCA components, and hyper-parameters were kept equal throughout all tests.

Table 7. Logistic Regression Testing - Comparison

Distance Measure	Accuracy	AUC	Recall	Precision	F1
Full Song - DTW	0.5574	0.567	0.5652	0.2281	0.3250
Stem - DTW	0.6639	0.6653	0.5217	0.2857	0.3692
Full Song - CSS	0.7579	0.8865	0.7391	0.5000	0.5965
Stem - CSS	0.8607	0.8507	0.4783	0.6875	0.5641
Stem & Min(d) - CSS	0.8852	0.8863	0.4783	0.8462	0.6111
Stem, Min(d) & DTW Stem	0.9016	0.9029	0.5217	0.9231	0.6667

Given these results, we believe there is evidence of improvement in overall model performance through the judicial use of stemming. A combined approach using stem distance and overall distance measures achieved an increase of 61.8% increase in Accuracy, a 304.7% increase in Precision, a 59.2% increase in AUC, and a 105.1% increase in F1 score. Recall changed marginally for the combined

models. Given the nature of the music industry, Precision was utilized as our optimization metric for all models. Precision is more important than Recall for this case. It is much more important for an automated system to be sure that a song labeled as a cover song, or plagiarism in the extreme case, is actually supposed to be labeled that way.

As we have said, we believe that these methods can be adapted to detect plagiarism. As an experiment, we located 46 songs either confirmed as plagiarism through court case losses and compensatory damages, or settled out of court. We then utilized trained models for each of our methods to determine, at an aggregated sense, whether our models were able to detect potential plagiarist features within the song pairs. As this data was unseen by previous training, this acted as a secondary measure of performance for the end-state goal of detecting potential plagiarism.

Table 8. Detecting Plagiarism - Comparison

	% Detected
Full Song - DTW	0.0%
Stem - DTW	2.2%
Full Song - CSS	8.6%
Stem - CSS	8.6%
Stem & Min(d) - CSS	9.8%
Stem, Min(d) & DTW Stem	45.7%

Most models performed poorly when applied to potentially plagiarized songs. However, our final model that combined stemming for euclidean song similarity, Min(d) as a feature for the minimum distance of all euclidean distances, and dynamic time warping distance for stems, saw material results. The model containing those distance features was able to detect 45.7% of the songs in the deployment test set as having potentially plagiarized features within the songs. We view this as a meaningful step towards potentially building an automated song plagiarism detection system, by leveraging known cover songs for training. And, we believe these techniques are particularly useful given the industry widespread use of sampling and remixing.

7 Applicability & Future Work

For future work, we believe that source separation (Stemming) of each song pair to improve song similarity measurement should be considered as a baseline prior to feature extraction. Advancing our similarity technique from a whole song comparison to a partial song comparison can further expand the field, and address

the partial sampling and re-mixing aspect of music plagiarism. For example, an artist can sample five seconds of another song and embed that sample inside of a four minute creative work. Detecting that sample adds additional complexity and another dynamic in detecting music plagiarism through similarity distance measurements using Dynamic Time Warping. Another expansion in this field would be detecting the samples that are the inverse or backwards rendition of the original work.

We believe that our approach of stemming in our Song Comparison Preprocessing Pipeline can also be relevant to additional research areas such as recommendation and search; making search and recommendation use cases much more targeted in the sense of being able to isolate search to specific stemmed features. These approaches, coupled with applying these techniques to detect potential plagiarism in songs automatically, could significantly help artist and royalty owners in not only protecting copyright for artistic work, but also improving exposure and segmentation for those works.

8 Lessons Learned

Music audio analysis is a rich field, and given its roots in signal processing its not surprising at the wide array for analysis tools and measurements available to the field. That being said, the music industry overall is still reeling from the digitalization of its industry. This brings new challenges, and new opportunities. Given the proliferation of sampling, remixing, and outright copyright infringement that exists in the market, new tools are needed urgently.

Much of the research today focuses on recommendation and ranking. However tools for recognizing cover songs, and potential plagiarism, have an enormous opportunity. In the present music market, its the artist that loses, and the artist who has the least control over their work. Being able to detect musical work similarity is of the highest priority for that under-served, under-powered group.

It is for these reasons we first took up this project. We believe that through our analysis and work that there is real merit to modifying the preprocessing pipelines for many research avenues to include source separation (stemming), at least in rudimentary form. The problem with music sampling, and the theft of those samples, is that it can happen to any portion of a song. Source separation allows for a tighter window of comparison between two songs, thereby increasing the overall resolution of similarity measurements for songs. We have seen how this can impact classification performance by improving Accuracy and Precision markedly, and we demonstrate this with a simple, but effective, use of PCA and regularized logistic regression. We could easily see how stemming could be effectively applied with state of the art techniques in Deep Learning

pipelines leveraging source separated stems for MFCC and HPCP features and cross similarity measurements.

9 Conclusion

We believe we have shown the value of source separated content as it relates to classification for cover songs. This is directly applicable to cover song detection, as well as similarity ranking tasks, and song plagiarism domains. Additionally, we believe the practice could be easily woven into existing systems as a preprocessing step to generate track specific spectral and rhythmic features for songs. That being said, until the artists of the world are armed with tools to protect their craft, we believe they will be under-represented for their work. We encourage others to continue to develop these capabilities, not for the large companies, but for the individual artists that so desperately needs advocates.

References

1. Rentfrow, Peter J.; Goldberg, Lewis R.; Levitin, Daniel J. (1 January 2011). "The Structure of Musical Preferences: A Five-Factor Model" Journal of Personality and Social Psychology. 100 (6): 1139–1157.
2. https://ijtre.com/images/scripts/2020071023.pdf Siddaraj, M. G., et al. "COVER SONG DETECTION."
3. "How to Tell If a Song's Been Copied - from a Trained Musicologist - BBC Newsbeat." BBC News, BBC, 23 Sept. 2015, www.bbc.co.uk/newsbeat/article/34282895/how-to-tell-if-a-songs-been-copied-from-a-trained-musicologist.
4. International Federation of the Phonographic Industry. IFPI Global Music Report 2019, 2 Apr. 2019, ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2019.
5. Zhang, Bingjun, et al. "CompositeMap." Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 09, 2009, https://doi.org/10.1145/1571941.1572011.
6. Tao Feng. "Deep Learning for music genre classification"
7. Office, U.S. Copyright. "Legislative Developments." Copyright, www.copyright.gov/legislation/dmca.pdf.
8. Muller, M. "Fundamentals of Music Processing: Audio, analysis, algorithms, applications" SPRINGER. (2016)
9. Essentia. (n.d.). Retrieved June 18, 2020, from https://essentia.upf.edu/
10. Smith–Waterman algorithm. (2020, March 08). Retrieved June 18, 2020, from https://en.wikipedia.org/wiki/Smith
11. Ren, Zhen, Chunxiao Fan, and Yue Ming. "Music retrieval based on rhythm content and dynamic time warping method." 2016 IEEE 13th International Conference on Signal Processing (ICSP). IEEE, 2016.
12. Novak, David, Petr Volny, and Pavel Zezula. "Generic Subsequence Matching Framework: Modularity, Flexibility, Efficiency." International Conference on Database and Expert Systems Applications. Springer, Berlin, Heidelberg, 2012.

13. Holzapfel, Andre, Bob Sturm, and Mark Coeckelbergh. "Ethical dimensions of music information retrieval technology." Transactions of the International Society for Music Information Retrieval 1.1 (2018): 44-55
14. Hardjono, Thomas, et al. "Towards an Open and Scalable Music Metadata Layer." arXiv preprint arXiv:1911.08278 (2019).
15. D. Bogdanov, J. Serra, N. Wack, P. Herrera and X. Serra, "Unifying Low-Level and High-Level Music Similarity Measures," in IEEE Transactions on Multimedia, vol. 13, no. 4, pp. 687-701, Aug. 2011.
16. Tao Li and M. Ogihara, "Content-based music similarity search and emotion detection," 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Que., 2004, pp. V-705.
17. Logan, Beth, and Ariel Salomon. "A content-based music similarity function." Cambridge Research Labs-Tech Report(2001).
18. Dannenberg, Roger B., Belinda Thom, and David Watson. "A machine learning approach to musical style recognition." (1997).
19. Kuo, Fang-Fei, et al. "Emotion-based music recommendation by association discovery from film music." Proceedings of the 13th annual ACM international conference on Multimedia. 2005.
20. Tralie, C. J. (2017). Early mfcc and hpcp fusion for robust cover song identification. arXiv preprint arXiv:1707.04680.
21. Chang, S., Lee, J., Choe, S. K., Lee, K. (2017). Audio cover song identification using convolutional neural network. arXiv preprint arXiv:1712.00166.
22. Lee, J., Chang, S., Choe, S. K., Lee, K. (2018, April). Cover Song Identification Using Song-to-Song Cross-Similarity Matrix with Convolutional Neural Network. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 396-400). IEEE.
23. Ellis, D. (n.d.). Retrieved Aug 18, 2019, from https://labrosa.ee.columbia.edu/projects/congs/c80/
24. Dittmar, Christian, et al. "Audio forensics meets music information retrieval—a toolbox for inspection of music plagiarism." 2012 Proceedings of the 20th European signal processing conference (EUSIPCO). IEEE, 2012.
25. Deezer. (2020, June 18). Deezer/spleeter. Retrieved June 18, 2020, from https://github.com/deezer/spleeter